WATER RESOURCES RESEARCH, VOL. 44, W02437, doi:10.1029/2007WR006104, 2008

Multivariate streamflow forecasting using independent component analysis

Seth Westra,[1,3] Ashish Sharma,[1] Casey Brown,[2] and Upmanu Lall[2]

Received 10 April 2007; revised 26 September 2007; accepted 9 November 2007; published 27 February 2008.

[1] Seasonal forecasting of streamflow provides many benefits to society, by improving our ability to plan and adapt to changing water supplies. A common approach to developing these forecasts is to use statistical methods that link a set of predictors representing climate state as it relates to historical streamflow, and then using this model to project streamflow one or more seasons in advance based on current or a projected climate state. We present an approach for forecasting multivariate time series using independent component analysis (ICA) to transform the multivariate data to a set of univariate time series that are mutually independent, thereby allowing for the much broader class of univariate models to provide seasonal forecasts for each transformed series. Uncertainty is incorporated by bootstrapping the error component of each univariate model so that the probability distribution of the errors is maintained. Although all analyses are performed on univariate time series, the spatial dependence of the streamflow is captured by applying the inverse ICA transform to the predicted univariate series. We demonstrate the technique on a multivariate streamflow data set in Colombia, South America, by comparing the results to a range of other commonly used forecasting methods. The results show that the ICA-based technique is significantly better at representing spatial dependence, while not resulting in any loss of ability in capturing temporal dependence. As such, the ICA-based technique would be expected to yield considerable advantages when used in a probabilistic setting to manage large reservoir systems with multiple inflows or data collection points.

Citation: Westra, S., A. Sharma, C. Brown, and U. Lall (2008), Multivariate streamflow forecasting using independent component analysis, *Water Resour. Res.*, *44*, W02437, doi:10.1029/2007WR006104.

1. Introduction

[2] Providing seasonal forecasts of rainfall and/or streamflow is an important challenge in hydrology, with potential benefits in reservoir management, operation of irrigation networks, and flood control, among others. These forecasts can be particularly pertinent for regions that experience significant inter-annual variability that result from the El Niño Southern Oscillation (ENSO) phenomenon [*Chiew and McMahon*, 2002; *Sharma*, 2000a], as it is often possible to use knowledge of ENSO and related oceanic patterns to provide estimates of future rainfall and/or streamflow that outperform the climatological means. Although societal benefits of these forecasts can be difficult to quantify, several studies have been published recently showing significant economic gains through the application of such forecasts to hydropower generation [e.g., *Hamlet et al.*, 2002; *Yao and Georgakakos*, 2001].

[3] These forecasts are commonly classified as either statistical or dynamical, with an excellent review provided by *Goddard et al.* [2001]. The present paper concerns itself with the field of statistical forecasts, which involves identifying mathematical relationships between a set of predictors such as global sea surface temperatures (SSTs), and the variables one wishes to predict (predictands), which for water resources applications typically includes rainfall and/or streamflow at a catchment or regional scale. To be considered successful, these forecasts generally are expected at minimum to outperform the climatological mean, as well as predictions obtained by adopting a simple persistence model that assumes recent climate anomalies will continue into the following season [*Huang et al.*, 1996].

[4] A review of a large number of statistical forecast models around the world [*Goddard et al.*, 2001] suggests that most of the predictability from these models is related to variability in the tropical Pacific, and in particular in relation to the ENSO phenomenon [for example, see *Barnston*, 1994; *Casey*, 1998; *Chiew and McMahon*, 2002; *Filho and Lall*, 2003]. Thus it has become common to use indices of ENSO as the basis of the statistical model. Popular forms of such models include either parametric [e.g., *McBride and Nicholls*, 1983; *Singhrattna et al.*, 2005; *Woolddridge et al.*, 1999] or non-parametric [e.g., *Sharma*, 2000b; *Singhrattna et al.*, 2005] regression models, conditional probability models (by defining a given year as El Niño, La Niña or neutral, and estimating the conditional probability density function of the predictand according to this classifi-

[1]School of Civil and Environmental Engineering, The University of New South Wales, Sydney, NSW, Australia.
[2]Department of Earth and Environmental Engineering, Columbia University, New York, New York, USA.
[3]Now at Sinclair Knight Merz, NSW, Australia.

0043-1397/08/2007WR006104$09.00

cation; see *Mason and Goddard*, 2001; by defining a given year as El Niño, La Niña or neutral, and estimating the conditional probability density function of the predictand according to this classification; see *Ropelewski and Halpert*, 1996), and models that incorporate both ENSO state as well as a recent trend component [*Stone et al.*, 1996].

[5] In numerous instances, one may wish to consider climate features which are not adequately described by one of the ENSO indices as potential predictors to a statistical model [e.g., *Verdon et al.*, 2004]. These additional features may be described by other indices, such as the North Atlantic Oscillation [NAO; *Hurrell*, 1995; *Hurrell and Van Loon*, 1997] or the Indian Ocean Dipole [IOD; *Ashok et al.*, 2003; IOD; *Saji et al.*, 1999; *Saji and Yamagata*, 2003], however these indices represent climatological phenomena which are themselves often highly correlated with ENSO. An alternative to the index-based approach makes uses information contained in the large multivariate data sets such as global sea surface temperatures (SSTs), which has been shown to enhance the quality of forecasts in numerous instances [e.g., *Drosdowsky and Chambers*, 2001; *Nicholls*, 1989].

[6] As these global data fields are usually quite large, and contain significant spatial correlation, some form of dimension reduction is required prior to incorporation into a statistical model. In simple cases, this might involve application of principal components analysis (PCA) to the multivariate predictor data set to extract individual ‘modes’ of variability that are mutually uncorrelated and successively explain the maximum amount of remaining variance in the data [*Barnston and Ropelewski*, 1992]. This is frequently followed by applying a rotation to the PCA solution to enhance interpretability of the climate ‘modes’, with Varimax [*Richman*, 1986] and Independent Component Analysis [ICA; *Aires et al.*, 2000] being two popular choices. It is then usually a straightforward exercise to find a parametric or nonparametric statistical model to relate the predictors and predictand as the basis for generating the forecasts.

[7] This situation is more complicated when the predictand is also multivariate, such as when the aim is to forecast rainfall or streamflow in a large reservoir system where there are gauging stations at multiple sites. Arguably the most commonly used statistical technique for these data sets is canonical correlation analysis [CCA; *Barnett and Preisendorfer*, 1987; *Barnston*, 1994; *Barnston and Ropelewski*, 1992; *Hwang et al.*, 2001; *Nicholls*, 1989; *Shabbar and Barnston*, 1996; CCA; *Storch and Zwiers*, 2001]. The basic approach of CCA is to develop a linear relationship between a multivariate predictor set and a multivariate predictand set such that the sum of squared errors is minimized. This is achieved by performing an eigen-analysis on the cross-correlation matrix constructed by computing correlation coefficients between the predictor and predictand, such that the correlation explained between these data sets is maximized, while at the same time ensuring successive canonical variates are mutually uncorrelated.

[8] CCA is a very powerful multivariate method that has been used to develop numerous forecasts of considerable skill. One limitation, however, is that the model linking the predictor and predictand data sets is constrained to be linear, which may not reflect accurately the true relationship between these two data sets. Furthermore, by focusing on correlation statistics, CCA may ignore higher-order dependencies, which may reduce the performance of the approach in certain hydrological applications [see *Westra et al.*, 2007].

[9] This paper extends the work of *Westra et al.* [2007] to forecast multivariate streamflow time series in a catchment located in Colombia, South America. The approach commences by using ICA to transform the predictand data to a set of univariate components which are as independent from each other as possible. This allows the development of a model that is univariate in the predictand and multivariate in the predictors, where in this case the predictors are derived by applying PCA to the SST data set for dimension reduction followed by an ICA rotation. For simplicity we then use a linear model to link the ICs of the predictor data set to individual ICs of the predictand data set, although non-linear or non-parametric extensions are also possible. Finally, we apply the inverse ICA rotation to the estimated predictand ICs, such that the spatial dependence structure is maintained.

[10] This method is expected to yield a number of advantages compared with other multivariate methods. For example, although the selection of predictors does not rely on information contained within the predictand data set, as is the case with CCA, it can be argued that the additional flexibility gained by allowing rotations of the SST data set to enhance interpretability may increase the robustness of the ensuing model, as well as potentially assisting in understanding the drivers of climate within a particular region. Furthermore, the proposed approach incorporates greater flexibility both in terms of selecting the predictor data set (e.g., using PCA, Varimax or ICA), as well as defining the relationship between predictor and predictand, which for the purposes of this research is represented by a linear model but can be easily extended to a range of non-linear or non-parametric models. Finally, considering independence in the predictand data set should ensure that the spatial dependence is better maintained compared to correlation-based methods.

[11] The remainder of the paper is as follows. The background to ICA is presented in section 2. This is followed by an overview of the streamflow and sea surface temperature data sets used in the analysis in section 3. The forecasting methodology is then presented in section 4, followed by the results in section 5. Finally, the conclusions from this study are presented in section 6.

2. Independent Component Analysis

[12] Independent Component Analysis (ICA) is a recently developed mathematical technique which is used to separate mixtures of signals by maximizing the independence of the extracted components [*Comon*, 1994; *Herault and Jutten*, 1986]. The primary motivation for ICA traditionally has been the link between finding independent representations of multivariate data and solving the blind source separation (BSS) problem, in which one wishes to derive a set of independent ‘source signals’ from a set of observations, having no information about either the nature of these signals, or the manner in which they have been mixed [*Hyvarinen*, 1999; *Hyvarinen et al.*, 2001; *Lee*, 1998; *Oja*, 2004]. This has been the main justification for applying ICA to sea surface temperatures [SSTs; *Aires et al.*, 1999; *Aires et al.*, 2000; *Aires et al.*, 2002; *Basak et al.*, 2004; *Ilin et al.*,

2006] since it allows the extraction of dominant 'modes', or 'features', of the underlying system.

[13] An additional benefit to maximizing independence is that it allows multivariate data to be decomposed into a set of univariate series, with each series exhibiting minimal dependence on the other series. In certain cases this can simplify statistical models significantly, since each univariate series can be considered in isolation rather than as part of a complex multivariate model (refer to discussion by *Westra et al.*, 2007). Both the statistical and analytical advantages of maximizing independence will be exploited in this paper.

[14] The simplest form of ICA occurs when an m by l observation matrix $\mathbf{X}$ is derived through the mixing of an n by l 'source' matrix $\mathbf{S}$, which are commonly referred to as the independent components [*Comon*, 1994]. These ICs are assumed to be non-Gaussian (with the possible exception of at most one IC, since by knowing all but one IC, the final IC can be specified automatically), mutually statistically independent and zero-mean. In addition, it is assumed that $n \leq m$. Put into vector-matrix notation, and assuming that the mixing is both linear and stationary, yields:

$$\mathbf{X} = \mathbf{AS} \tag{1}$$

where $\mathbf{A}$ is known as the mixing matrix of dimension $m \times n$. The objective of ICA is to estimate the mixing matrix, $\mathbf{A}$, as well as the independent components, $\mathbf{S}$, knowing only the observations $\mathbf{X}$. This can be achieved up to some scalar multiple of $\mathbf{S}$, since any constant multiplying an independent component in equation (1) can be canceled by dividing the corresponding column of the mixing matrix $\mathbf{A}$ by the same constant.

[15] Central to the identification of the ICs from the data $\mathbf{X}$ is the assumption that all except at most one IC will be "maximally non-Gaussian" [*Hyvarinen et al.*, 2001]. This follows from the logic outlined in the central limit theorem, which is that if one mixes independent random variables through a linear transformation, the result will be a set of variables that tend to be Gaussian. If one reverses this logic, it can be presumed that the original independent components must have a distribution that has minimal similarity to a Gaussian distribution. Consequently, the approach adopted to extract ICs from data containing mixed signals amounts to finding a transformation that results in variables that exhibit maximal non-Gaussianity as defined through an appropriately specified statistic.

[16] ICA is frequently compared with a related technique known as principal component analysis (PCA), except that ICA results in components that are statistically independent, whereas PCA leads to components that, while being uncorrelated, may exhibit strong dependence on each other. The independent components are extracted using higher-order moment information, i.e., information other than that contained in the covariance matrix of $\mathbf{X}$ [*Oja*, 2004]. PCA is commonly used as a pre-processing step, however, both as a means of dimension reduction, and as a starting point for whitening (or sphering) the data such that $\mathbf{X}$ is linearly transformed into another n-dimensional vector $\mathbf{Z}$ that has a unit covariance matrix. For more detail on the statistical differences between uncorrelatedness and independence, refer to the analysis provided by *Westra et al.* [2007].

[17] In this study the ICA logic is applied to develop independent components of SSTs, which are used to formulate predictive models of hydrological responses, which in this case are the independent components derived from a vector of streamflows at key locations in the selected study area. More details on the study area and the data used are presented next.

3. Data

3.1. Streamflow

[18] The streamflow data set that is used in this study consists of 20 stations located in the Magdalena-Cauca in Colombia, and which drains to the Caribbean Sea. The locations of these stations are shown in Figure 1, with each station located at a reservoir inflow point representing observed inflows to three hydroelectric generating reservoirs. Each station retains the name of the river on which it is located, and measurements were initiated during the design phase of each proposed reservoir/dam and continue to the present. The data represents unimpaired monthly total flows from 1963 to 2004, which in certain cases (e.g., the Alto Anchicaya from 1963–1975) has been calibrated using regions with longer and more reliable records. These monthly flows were aggregated into seasonal flows, defined as summer (DJF), autumn (MAM), winter (JJA), and spring (SON).

[19] The objective of this study is to outline the use of independent component analysis to transform the multivariate data into a set of univariate components, so that univariate regression methods can be applied to generate forecasts. Because of the relatively short duration of the data (42 data points), it is not possible to apply ICA to the full 20 dimensional data set, and as such we consider two trivariate subsets of this data for further analysis.

[20] The first subset is derived using a K-medoid clustering algorithm [*Hastie et al.*, 2001] with K = 3 to the streamflow data to find the three 'clusters' which maximize the pair-wise dissimilarity between those sites assigned to the same cluster and those in different clusters. The sites selected are those which are closest to the center of each cluster. This approach was preferred over K-means clustering, since data from a 'real' streamflow station could be used for subsequent analysis rather than the mathematical centroid of each cluster which by averaging may no longer have the characteristics of a true streamflow time series. The cluster analysis was performed on the annual streamflow data, rather than on seasonal data, to simplify ensuing calculations by ensuring that the sites used for the analysis were the same for each season. The selected sites were Guavio, Guatape and Rio Grande.

[21] The second subset consists of the same streamflow stations that were used to illustrate the difference between uncorrelated and independent components in the context of stochastic analysis provided by *Westra et al.* [2007], and includes Rio Grande, Salvajina and Alto Anchicaya. The data from these stations exhibit significantly more spatial dependence than the data from the clustering analysis described above, and therefore allows a more rigorous comparison of the ability of a range of statistical techniques in maintaining this spatial dependence in the context of streamflow forecasting. The gauging stations used in the

Figure 1. Streamflow stations used in the analysis.

analysis, including seasonal and annual mean streamflows at each stations, are provided in Table 1.

3.2. Climate

[22] A global sea surface temperature anomaly (SSTA) data set was obtained from a reconstruction of raw SST values using an optimal smoother, as described in [*Kaplan et al.*, 1998]. On the basis of a simple correlation analysis between individual streamflow time series and the global data set (results not shown), it was found that statistically significant correlations were found for certain stations and certain seasons in all major oceans and for the full range of latitudes. As a result, we decided not to reduce the global data set to a smaller geographic range. The linear trend of about 0.6 degrees Celsius was removed from each SSTA time series before continuing with the analysis.

[23] An index of the oceanic component of the ENSO phenomenon, Niño 3.4, was also used in this analysis, and is defined as the seasonally averaged SSTA over the eastern Pacific [5°S – 5°N, 120°W – 170°W; *Trenberth*, 1997]. Both the global SSTA data set and the Niño 3.4 index were obtained from the International Research Institute for Climate and Society (IRI) website (http://iri.columbia.edu). In each case, the data was aggregated to form seasonal time

Table 1. List of Gauging Stations Used in the Analysis, Together With Seasonal and Annual Mean Streamflow (m^3/s)[a]

Station	Record	DJF	MAM	JJA	SON	Annual
Guavio	1963–2004	70.2 (22.9)	201.7 (53.77)	399.5 (75.6)	192.8 (32.5)	**864 (121)**
Guatape	1959–2004	74.4 (22.3)	97.9 (26.8)	88.9 (21.0)	130.9 (27.4)	**392 (74)**
Rio Grande	1942–2004	78.7 (22.9)	95.8 (27.6)	100.9 (26.7)	126.6 (30.0)	**402 (85)**
Salvajina	1947–2004	512.2 (194.6)	436.1 (126.7)	298.6 (51.1)	370.3 (115.3)	**1617 (357)**
Alto Anchicaya	1963–2004	130.5 (29.5)	139.8 (28.6)	105.6 (26.8)	160.9 (28.4)	**537 (78)**

[a]Standard deviations are presented in parentheses.

series. For the subsequent statistical model, concurrent data was used throughout the analysis.

4. Methodology

4.1. Forecasting Overview

[24] The objective of this paper is to describe various methods to forecast seasonal streamflow for a water supply catchment in Colombia, and demonstrate certain benefits in formulating the forecasting models by considering independence in both predictor and predictand data sets. Each of the forecasting approaches will be applied to both subsets of three streamflow stations identified in section 3.

[25] Let the trivariate streamflow data be represented by $\mathbf{X}_{q,t} = \{\mathbf{x}_{q,1,t}, \mathbf{x}_{q,2,t}, \mathbf{x}_{q,3,t}\}$. The subscript, q, will be used throughout this paper to indicate a streamflow variable, whereas the subscript, s, will be used to refer to the SST data set. All streamflow data has been normalized by subtracting the seasonal mean and dividing by the standard deviation, and as such represents seasonal anomaly data. Note that each $\mathbf{x}_{q,i,t}$ represents a time series for a particular season, with the season denoted by t. Arguably the most simple forecast model is to consider the previous season streamflow time series (t-1) in an autoregressive order-1 model, as described below:

$$\mathbf{x}_{q,i,t} = \beta_{AR,t}\mathbf{x}_{q,t-1} + \boldsymbol{\varepsilon}_{i,t}^{(1)} \tag{2}$$

where $\beta_{AR,t}$ represents the regression coefficient for that season for the autoregressive model, and $i \in \{1, 2, 3\}$ represents individual streamflow time series, such that a separate model is developed for each streamflow station. The previous season streamflow, represented by $\mathbf{x}_{q,t-1}$ is taken as the unweighted average streamflow from each of the three stations, and therefore does not contain the subscript i. This was preferred over using a separate predictor (i.e., streamflow for the previous season at the same site) for each model, in the interest of keeping the modeling structure as parsimonious as possible, and also because of the spatial linkages the respective flows have with each other.

[26] The approach represented by equation 2 is equivalent to the standard climatology plus persistence model, and as discussed in the Introduction of this paper, any forecast model should be tested against this basic model to determine whether using exogenous variables relating to a climate state will result in an increase in predictability. As a result, the remainder of this paper will focus on improving this basic model by focusing on minimizing the error component, $\varepsilon_{i,t}^{(1)}$. Four such approaches will be considered, including:

[27] 1) Considering an index of the ENSO phenomenon, Niño 3.4, as the predictor, and regressing this predictor against the error $\varepsilon_{i,t}^{(1)}$ associated with each univariate streamflow time series individually;

[28] 2) Using ICA on the SSTA data to obtain a multivariate SSTA predictor field, which will be regressed against each univariate streamflow error time series $\varepsilon_{i,t}^{(1)}$ individually;

[29] 3) Applying CCA to model both multivariate predictor and multivariate predictand, by using eigen-decomposition techniques to maintain second order dependencies in the data; and

[30] 3) Applying ICA to both the multivariate predictor and multivariate predictand, using an information-theoretic approach to maintain second- and higher-order dependencies in the data.

[31] Note that both the Niño 3.4 index and the SSTA data used for the forecasts are concurrent, to simplify the analysis. In reality, of course, the user will not have access to concurrent information to develop future projections, and as such, this method will need to be modified to incorporate a forecasting model for SSTAs (using either statistical or dynamical means), or alternatively using a lagged relationship between predictor and predictand.

4.2. Approach 1 - Regression of the Niño 3.4 Index Against Univariate Streamflow Residuals

[32] We present first a simple regression model that uses the Niño 3.4 index as a measure of the oceanic component of ENSO variability:

$$\mathbf{x}_{q,i,t} = \beta_{AR}\mathbf{x}_{q,t-1} + \beta_{NINO34}\mathbf{z}_t + \boldsymbol{\varepsilon}_{i,t}^{(2)} \tag{3}$$

where $\mathbf{z}_t$ represents the seasonal Niño 3.4 index at season t, and β_{NINO34} represents the regression coefficient for this predictor. This model is mathematically and computationally simple, and is frequently also highly interpretable, since a statistically significant model will demonstrate a direct relationship between the hydrological variable of interest, and a known measure of exogenous climate variability. Its main disadvantages is that each streamflow time series is modeled independently, and hence cannot be expected to exhibit the spatial dependence one would observe in the historical data. An additional disadvantage is the reliance on ENSO as the only mode of variability, a strong assumption given evidence of secondary effects that change streamflow patterns around the world.

4.3. Approach 2 - Regression of Multivariate SSTAs Against Univariate Streamflow Residuals

[33] Our second model includes sea surface temperature anomalies as an additional factor in predicting flows, consistent with the recommendations of [*Paegle and Mo*, 2002]. This is done by first applying PCA to reduce the dimension of the global SSTA data set, followed by application of ICA to the retained components to enhance the interpretability of each component. We considered six components in total, which is based on a compromise between maximizing interpretability of individual ICs while at the same time maximizing the variance explained in the original data set. This allows two predictors on average for each of the three predictands. The ensuing model is written as follows:

$$\mathbf{x}_{q,i,t} = \beta_{AR}\mathbf{x}_{q,t-1} + \sum_j \beta_j \mathbf{y}_{s,j,t} + \boldsymbol{\varepsilon}_{i,t}^{(3)} \tag{4}$$

where $\mathbf{y}_{s,j,t}$ represents estimates of the independent components, $\mathbf{S}$, obtained by applying equation (1) to the SSTA data set for season t. Here, $j \in \{1, 2, \ldots, 6\}$ represents the subset of predictors obtained from the set of independent components from the SSTA data set, selected using a forward stepwise selection procedure with the partial t-statistic for testing whether a predictor should be included. We use a 90 percentile cut-off t value of 1.68 which corresponds to a correlation coefficient of 0.255. It was found that a large number of predictors were excluded using the slightly higher 95 percentile t value, which we believe is partly due to the difficultly in demonstrating statistical significance for such short data sets, particularly when looking beyond the first predictor, which is usually related to the ENSO phenomenon. It should be noted that this model would be expected to reflect variability due to a broader set of causes than just ENSO, but would still be constrained in its representation of spatial dependence across the streamflow variables being modeled.

4.4. Approach 3 - Use of Canonical Correlation Analysis to Relate the Multivariate SSTAs to Multivariate Streamflow

[34] The disadvantage of the above two methods in the present context is that a separate model is formulated for each of the streamflow time series, without any provision to ensure an accurate representation of the spatial dependence that exists across the streamflow variable vector. A widely used approach for incorporating spatial dependence is known as canonical correlation analysis (CCA), which is a mathematical approach for identifying pairs of patterns in two multivariate data sets, and constructing sets of transformed variables by projecting the original data onto these patterns [*Wilks*, 2006]. For the purposes of this study, the predictors and predictands are the global SST data set represented by $\mathbf{X}_{s,t}$, and the trivariate reservoir inflow data set after accounting for persistence represented by $\mathbf{E}_t^{(1)}$, respectively.

[35] The following is a brief overview of CCA, based on the discussion by *Wilks* [2006] but with notation adjusted to be consistent with notation used elsewhere in this paper.

[36] The aim of CCA is to transform the original (centered) multivariate data sets $\mathbf{X}_{s,t}$ and $\mathbf{E}_t^{(1)}$ into a set of canonical variates, $\mathbf{V}$ and $\mathbf{W}$, defined by:

$$\mathbf{V} = \mathbf{A}^T\mathbf{X}_{s,t} \tag{5}$$

$$\mathbf{W} = \mathbf{B}^T\mathbf{E}_t^{(1)} \tag{6}$$

The vectors of weights, $\mathbf{A}$ and $\mathbf{B}$, are the canonical vectors, which are obtained through an eigen-decomposition of $\mathbf{X}_{s,t}$ and $\mathbf{E}_t^{(1)}$, with further details provided by *Wilks* [2006]. This results in the canonical variates having the properties

$$Corr[\mathbf{V}_i, \mathbf{W}_n] = \begin{cases} r_{Ci}, i = n \\ 0, i \neq n \end{cases} \tag{7}$$

where r_{Ci} is obtained through an eigen decomposition of the joint predictand-predictor covaiances matrices. Since the r_{Ci} allows each element of $\mathbf{V}$ to be related to a unique element of $\mathbf{W}$, estimation of $\mathbf{W}$ can be accomplished through the following linear relationship:

$$\hat{\mathbf{W}} = [\mathbf{R}_C]\mathbf{V} \tag{8}$$

where

$$[\mathbf{R}_C] = \begin{bmatrix} r_{C1} & 0 & 0 \\ 0 & r_{C2} & 0 \\ 0 & 0 & r_{C3} \end{bmatrix} \tag{9}$$

Using $\hat{\mathbf{W}}$, it is possible to develop an estimate of $\hat{\mathbf{E}}_t^{(1)}$ through the inverse of equation (6), which completes the specification of the CCA based forecasting procedure used in our study.

4.5. Approach 4 - Regression of Multivariate SSTAs Against Multivariate Streamflow Residuals

[37] Finally, we present an alternative approach that follows on from the multivariate resampling logic presented by *Westra et al.* [2007], which is based on transforming the multivariate predictor and predictand into a set of univariate series which do not exhibit dependence on each other. The ICA-based approach was compared with a PCA-based approach by *Westra et al.* [2007], and it was found that by focusing on statistical independence rather than correlation, ICA better represented spatial dependence of the multivariate data.

[38] A similar advantage is expected in this paper, where CCA considers only the information contained in the covariance matrix (second order statistics), whereas ICA provides an additional rotation to maximize some measure of statistical dependence. The forecasting approach adopted in this paper is presented below, and further details on the statistical properties of ICA are given by *Hyvarinen et al.* [2001] and *Westra et al.* [2007].

[39] We start by transforming the error term in equation (2) to yield:

$$\mathbf{Y}_{q(\varepsilon),t} = \mathbf{W}_{q(\varepsilon),t}\mathbf{E}_t^{(1)} \tag{10}$$

This equation is analogous to the inverse form of equation (1), where $\mathbf{Y}_{q(\varepsilon)t}$ represents an estimate of a set of independent components, $\mathbf{E}_t^{(1)}$ represents the original mixed multidimensional error matrix from equation (2) and $\mathbf{W}$ represents an estimate of the inverse of the mixing matrix, $\mathbf{A}$. Note we use bold to represent the full multivariate data, and as such the subscript i is not included.

[40] Since $\mathbf{Y}_{q(\varepsilon),t}$ consists of a set of independent components, it can be factorized into $\mathbf{y}_{q(\varepsilon),i,t}$, with $i \in \{1, 2, 3\}$, without loss of information. This allows for individual $\mathbf{y}_{q(\varepsilon),i,t}$ to be related to selected independent components of the multivariate SSTA data as was done in the previous approach. This gives us the following model:

$$\mathbf{y}_{q(\varepsilon),i,t} = \sum_j \beta_j \mathbf{y}_{s,j,t} + \boldsymbol{\varepsilon}_{i,t}^{(4)} \tag{11}$$

with each variable defined as before. Because the $\mathbf{y}_{q(\varepsilon),i,t}$'s are mutually independent, it is possible to generate the forecasting model without distorting the spatial dependence structure inherent in the original data. To achieve this, however, a modification to the forward selection process used in section 4.3 is necessary, to ensure that different $\mathbf{y}_{q(\varepsilon),i,t}$'s do not end up with the same predictor. This is achieved as follows:

[41] 1) Start by sorting the $\mathbf{y}_{q(\varepsilon),i,t}$'s by variance explained, with $\mathbf{y}_{q(\varepsilon),1,t}$ representing the maximum variance, and $\mathbf{y}_{q(\varepsilon),3,t}$ representing the minimum variance;

[42] 2) For $\mathbf{y}_{q(\varepsilon),1,t}$, find the SST independent component $\mathbf{y}_{s,j,t}$ that yields the maximum correlation coefficient. If this is higher than the 90 percentile t statistic cut-off of 1.68, retain this as a predictor of $\mathbf{y}_{q(\varepsilon),1,t}$, and remove from the pool of available predictors;

Table 2. Correlation Coefficients (r) From a Simple Persistence Model Together With Four Alternative Statistical Models That Use Some Measure of Climate State (Either the Niño 3.4 Index or a Dimension-Reduced Representation of Global Sea Surface Temperature Anomalies) to Provide Estimates of Streamflow at Three Separate Locations in Colombia[a]

Season	Station	Persistence	Persistence + Niño 3.4	Persistence + SST ICs - Univariate	Persistence + SST (CCA) - Multivariate	Persistence + SST ICs - Multivariate
Subset 1- results for stations - (1) Guavio, (2) Guatape and (3) Rio Grande						
DJF	1	0.40	0.33	0.40	0.24	0.40
DJF	2	0.38	0.39	0.50	0.50	0.46
DJF	3	0.67	0.77	0.76	0.76	0.75
MAM	1	0.46	0.35	0.49	0.41	0.47
MAM	2	0.50	0.48	0.53	0.42	0.55
MAM	3	0.52	0.53	0.61	0.56	0.53
JJA	1	0.29	0.24	0.29	0.15	0.23
JJA	2	0.06	0.30	0.56	0.46	0.39
JJA	3	0.40	0.64	0.67	0.68	0.54
SON	1	0.06	0.14	0.29	0.11	0.26
SON	2	0.15	0.17	0.38	0.30	0.35
SON	3	0.23	0.66	0.66	0.70	0.62
Subset 2 - results for stations - (1) Rio Grande, (2) Salvajina and (3) Alto Anchicaya						
DJF	1	0.67	0.74	0.73	0.74	0.71
DJF	2	0.67	0.73	0.69	0.71	0.70
DJF	3	0.55	0.69	0.71	0.69	0.64
MAM	1	0.65	0.62	0.65	0.61	0.65
MAM	2	0.74	0.70	0.74	0.68	0.74
MAM	3	0.37	0.25	0.37	0.24	0.37
JJA	1	0.56	0.72	0.71	0.74	0.72
JJA	2	0.44	0.55	0.50	0.50	0.52
JJA	3	0.23	0.56	0.65	0.64	0.65
SON	1	0.58	0.66	0.62	0.68	0.68
SON	2	0.54	0.60	0.58	0.58	0.59
SON	3	0.41	0.46	0.53	0.45	0.48

[a]All results presented have been cross validated using leave-one-out cross validation.

[43] 3) Repeat step (2) for $\mathbf{y}_{q(\varepsilon),2,t}$ and $\mathbf{y}_{q(\varepsilon),3,t}$, so that each $\mathbf{y}_{q(\varepsilon),i,t}$ has at most one predictor;

[44] 4) Repeat steps (2) and (3) to add a second predictor for each case where the t statistic is higher than the cut-off value of 1.68. Each time a predictor is added to the model, it should be removed from the pool of available predictors so that a given predictor can not be selected twice.

[45] We now have a model for each streamflow IC, consisting of at most two predictors. Using this model, we can estimate $\hat{\mathbf{y}}_{q(\varepsilon),i,t}$ for all i, followed by a rotation into original space using the equation:

$$\hat{\boldsymbol{\varepsilon}}_t^{(1)} = \mathbf{W}_{q(\varepsilon),t}^{-1}\hat{\mathbf{y}}_{q(\varepsilon),t} \tag{12}$$

Where $\mathbf{W}^{-1}$ represents the inverse of $\mathbf{W}$.

[46] The final estimates of streamflow can now be written as:

$$\hat{\mathbf{x}}_{q,i,t} = \beta_{AR}\mathbf{x}_{q,t-1} + \hat{\mathbf{e}}_{i,t}^{(1)} \tag{13}$$

We now have four alternative forecasting approaches, as well as a baseline persistence model, which we wish to compare both in terms of forecast performance for individual streamflow stations, and the ability to capture the spatial dependence of the multivariate data. This comparison is the subject of the following section.

5. Results

5.1. Spearman's Correlation as a Measure of Model Accuracy

[47] The objective of a forecast model is to make projections of a given predictand forward in time. The way our models are formulated, we first model explicitly the persistence of the streamflow variables on an aggregate streamflow at the previous time step. Hence an assessment of the suitability of this model structure is necessary. Here we use the correlation coefficient (r) between the expected streamflow and the true streamflow over the historical time period as the measure of model accuracy. As this measure does not consider spatial dependence, the univariate models would be expected to perform similarly to multivariate models.

[48] The results for both subsets of three streamflow stations are provided in Table 2. The results show that much of the predictive power is actually derived from the simple persistence model, with the majority of stations showing correlations greater than the 95 percentile coefficient for statistical significance, which was calculated to be 0.3. The correlation coefficients are not consistent between the individual stations and seasons, however, with a range from 0.06 to 0.74. This large variation from one station to the next suggests different response times for individual watersheds, as well as seasonal differences. Note that, although the station Rio Grande is common to both subsets, the correlation coefficients for the persistence model differ slightly. This is because the predictor for the persistence model is the unweighted average streamflow of all three sites at the previous season for a particular subset, and therefore the predictor is not common to both subsets.

[49] When factoring in exogenous information on climate state, whether through the Niño 3.4 index, the CCA model or through an ICA analysis of SSTAs, some improvement generally can be seen over the persistence model. Interestingly, however, this improvement does not significantly dependent on the nature of the climate model used. These

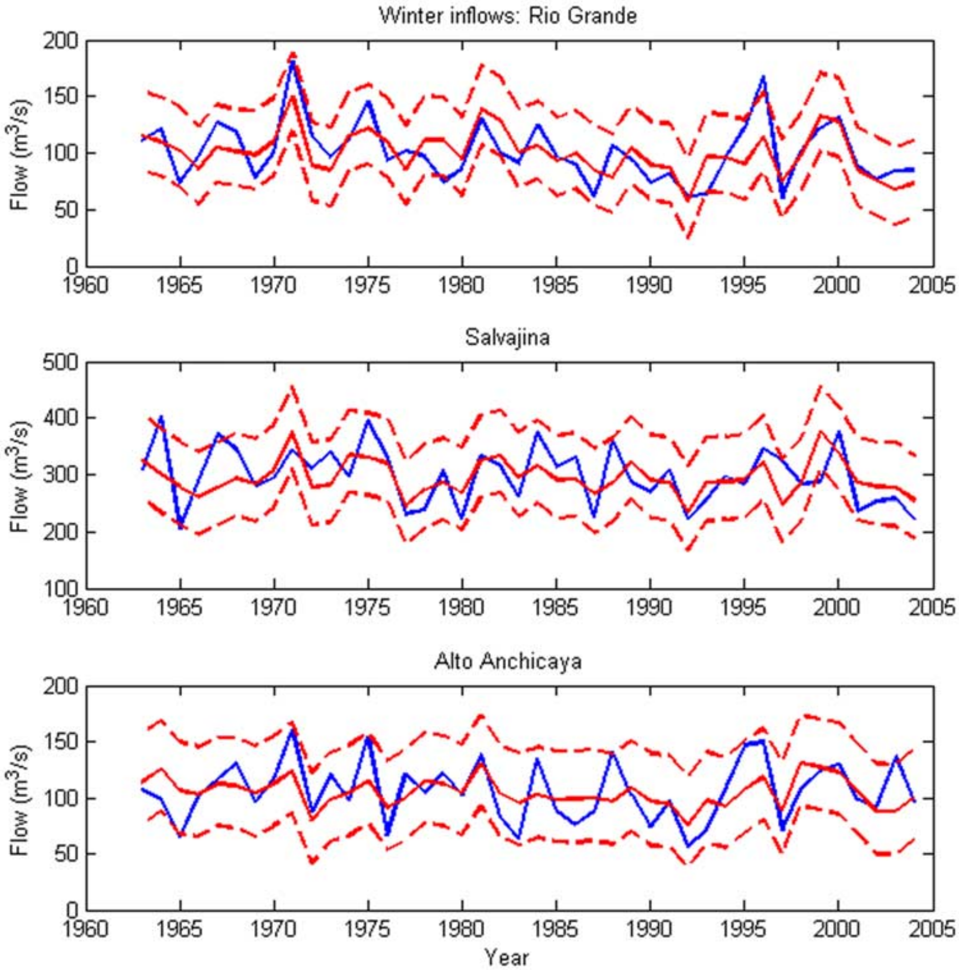

Figure 2. Time series of historical flows (blue line) and estimated flows based on multivariate ICA regression model (red solid line), for Rio Grande (top), Salvajina (middle) and Alto Anchicaya (bottom). Dotted red lines represent 5% and 95% significance levels.

results suggest that the streamflow time series analyzed here are largely persistence + ENSO driven, with little additional information derived from considering other regions of oceanic variability. Finally, it is noted that the different parameterizations of each modeling approach were accounted for by using leave-one-out cross validation in generating all the model results.

[50] To provide a visual assessment of the performance of the multivariate ICA model, the results of this model are compared with actual recorded winter inflows for each of the stations in Subset 2, and have been plotted in Figure 2. The recorded inflows are plotted as a blue line, with the best estimates from the multivariate ICA model presented as a solid red line. The correlation coefficients between recorded and estimated inflows for Rio Grande, Salvajina and Alto Anchicaya are 0.72, 0.52, and 0.65, respectively. The 5% and 95% confidence levels are plotted as dashed red lines, and are generated through bootstrapping the error terms for each approach and taking the 5th and 95th percentile for each estimate.

[51] The main result from this analysis is therefore that no significant improvement can be observed for the more complex persistence + SSTA ICs models compared with the more parsimonious persistence + Niño 3.4 model, so that from a univariate forecasting perspective there is a strong case in favor of choosing the simpler model to forecast streamflow. Similarly, however, there is no evidence of any loss in predictive power using the multivariate persistence + SSTA ICs model compared with the suite of other approaches analyzed, since this model performs equally well on a temporal basis compared to the univariate models. As such, the key question is how ICA represents the spatial dependence, and this is discussed further in the section below.

5.2. Spatial Dependence

[52] The main benefit of the multivariate ICA approach is its ability to capture the spatial dependence of the multivariate streamflow time series, and this can be explored only using a multivariate dependence measure. One such measure is the mean integrated squared bias (MISB), which is evaluated as follows [*Scott*, 1992]:

$$\int \left(\mathbf{f} - \hat{\mathbf{f}}\right) d\mathbf{X}_{q,t} \tag{14}$$

where **F** is the kernel density estimate [see *Sharma*, 2000b; *Westra et al.*, 2007, for details] of the original multivariate

Table 3. Mean Integrated Squared Bias (MISB) Results Comparing the Multivariate (Three Dimensional) Kernel Density Estimates of the Predicted Data Against the Three-Dimensional Kernel Density Estimates of the Original Data[a]

Season	Persistence	Persistence + Niño 3.4	Persistence + SST ICs - Univariate	Persistence + SST (CCA) - Multivariate	Persistence + SST ICs - Multivariate
Subset 1: Results for stations - Guavio, Guatape and Rio Grande					
DJF	0.0068	0.0069	0.0069	0.0061	0.0059
MAM	0.0056	0.0059	0.0058	0.0048	0.0042
JJA	0.0031	0.0032	0.0030	0.0022	0.0018
SON	0.0063	0.0059	0.0066	0.0037	0.0029
Subset 2: Results for stations - Rio Grande, Salvajina and Alto Anchicaya					
DJF	0.0700	0.0723	0.0713	0.0371	0.0299
MAM	0.0140	0.0162	0.0141	0.0089	0.0058
JJA	0.0097	0.0096	0.0071	0.0062	0.0051
SON	0.0159	0.0178	0.0179	0.0113	0.0094

[a]The MISB has been calculated for each streamflow model for four seasons, after cross-validation.

data, $\mathbf{X}_{q,t}$, which is taken to be the true density, and $\hat{\mathbf{f}}$ represents the kernel density estimate of the forecast data, $\hat{\mathbf{X}}_{q,t}$. The MISB is calculated using trivariate kernel density estimates to evaluate the joint dependence structure, and the results for each season and for each of the four models are provided in Table 3.

[53] Note that, to generate $\hat{\mathbf{X}}_{q,t}$ in this case we bootstrap the error terms in each approach to generate multiple plausible realizations of the forecasts. This enables the provision of forecasts in a probabilistic setting, so that the range of likely outcomes can be taken into account. The results in this section are based on 5000 such probabilistic forecasts of the streamflow variables being modeled.

[54] The results from Table 3 demonstrate that, in general, the univariate models (i.e., the persistence-only, the persistence + Niño 3.4 and the univariate persistence + SST ICs models) tend to perform similarly to each other, and have a higher MISB (i.e., poorer representation of spatial dependence) compared with the multivariate models. This difference is particularly notable for the second trivariate subset containing the stations Rio Grande, Salvajina and Alto Anchicaya, since the spatial correlation of the original data is greater than for the first trivariate subset. In contrast, the first subset, containing the stations Guavio, Guatape and Rio Grande, does not exhibit a great amount of spatial dependence, largely due to the clustering algorithm used to obtain the stations since this algorithm seeks stations that exhibit maximal within-cluster dependence while at the same time minimizing between-cluster dependence.

[55] Considering the multivariate models, it can be seen that although the CCA-based approach results in a significant improvement in the MISB score over the univariate approaches, the best results are reserved for the multivariate ICA approach, since this approach explicitly considers the full joint dependence in the data. The average percentage improvement in MISB score using the ICA-based approach for subsets 1 and 2 was 14% and 22%, respectively.

[56] As discussed earlier, the logic for the difference in results between the CCA- and ICA-based forecasting models is equivalent to the PCA- and ICA-based stochastic generation models compared by *Westra et al.* [2007]; that is, by considering the full dependence structure in a multivariate data set rather than focus on covariance or correlation-based statistics alone, it is possible to better simulate the joint probability density of the data. In this earlier paper, rather than consider two trivariate subsets, 992 trivariate combinations from the Colombia reservoir inflow data set were simulated, with the results showing an average improvement in the MISB score of 25% for the ICA-based approach. Although computational considerations preclude such a detailed analysis for the present study, the results for the two subsets considered here are consistent with these earlier results.

[57] Finally to illustrate the results from Table 3, we consider the performance of winter streamflow for the second subset, with joint dependence results presented in Figure 3. Only the CCA and ICA results are considered, since these are the only models that explicitly consider spatial dependence. The point to note is the smoothening in the multivariate results from both CCA and ICA in comparison to the raw data, which is to be expected because of the use of the bootstrapping procedure. While differences in the CCA and ICA results are not visually apparent, MISB calculations for the two indicate an improved representation of spatial dependence in the ICA model, as illustrated in Table 3.

6. Conclusions

[58] The objective of this paper is to demonstrate that generation of probabilistic multivariate seasonal streamflow forecasts using independent component analysis contains significant advantages over a range of alternative models that are commonly used. These alternative models include a simple persistence-only model, a persistence + Niño 3.4 model, a univariate persistence + SST ICs model and a multivariate CCA model. The results were evaluated both in terms of the ability to forecast seasonal streamflow, as well as whether the models were able to maintain the spatial variability that is present in the original data.

[59] When examining temporal dependence, with the exception of the persistence-only model, all other models perform comparably in terms of the correlation coefficient between true streamflow and estimated streamflow, suggesting that most of the variability in Colombia streamflow is a result of ENSO-driven processes. In contrast, a dramatic improvement can be seen between the multivariate ICA model compared with all the univariate models in terms of maintaining spatial dependence, and this improvement becomes more pronounced as the spatial proximity or climatological similarity between the stations is increased.

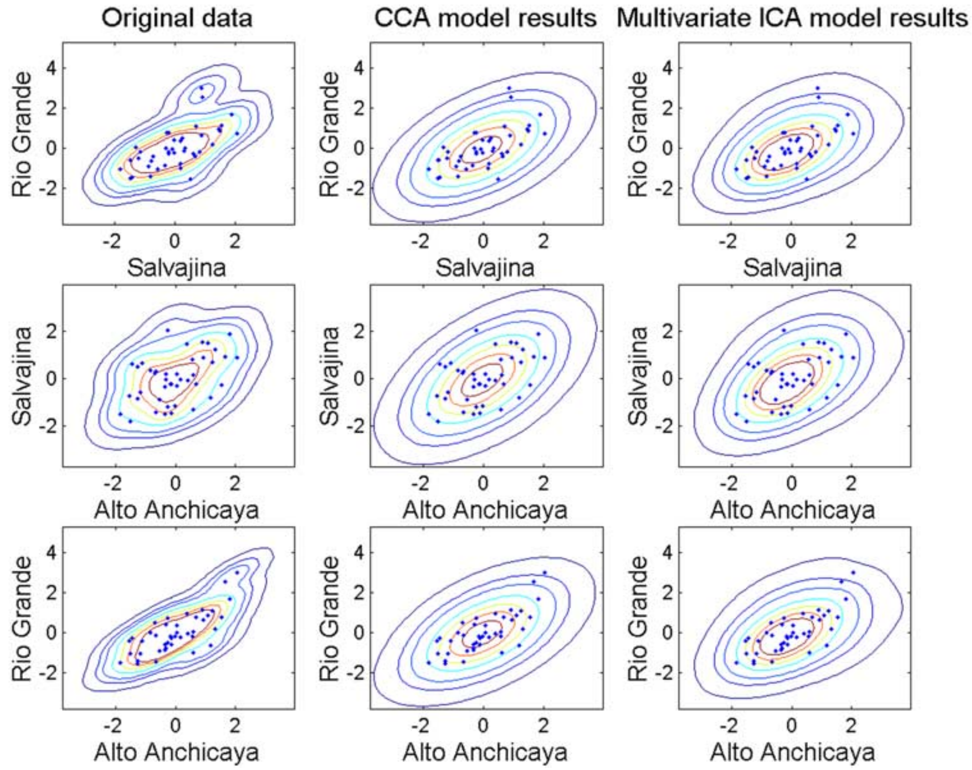

Figure 3. Spatial dependence between the winter reservoir inflows at the three stations in Subset 2 (Rio Grande, Salvajina and Alto Anchicaya) presented as bivariate plots. In each panel, the original (historical) data is represented as blue dots. The contours represent kernel density estimates of the original data (left panels), results from the CCA model (center panels) and the multivariate ICA model (right panels). The mean integrated squared bias (MISB) represents the difference in trivariate kernel density estimates of the original and modeled data, and was found to be 0.0062 for the CCA results, and 0.0051 for the ICA results.

[60] The multivariate ICA-based results also out-perform the CCA-based results in terms of spatial dependence, although by a lesser degree than compared with the univariate models. For the two subsets analyzed, the improvement in MISB using the multivariate ICA model was found to be 14% and 22% respectively. It was observed that the logic for this comparison is the same as the logic for comparing ICA with PCA by *Westra et al.* [2007], in that PCA and CCA both use second-order (correlation or covariance) statistics to generate orthogonal representations of the multivariate data. ICA also seeks an orthogonal representation of the data, but provides a further rotation to the data set to maximize dependence, defined by some higher-order variable such as skewness or kurtosis.

[61] A further advantage of the ICA-based method over CCA is that the structure of the multivariate ICA forecasting model is not fixed. In the present analysis we used a linear regression method to link predictors (SST ICs) with the predictands (streamflow ICs), however in certain cases it may be fruitful to pursue non-linear or non-parametric models instead for each IC response. Furthermore, although we used ICA to obtain the predictor data set, this may be changed to another rotational method such as Varimax PCA without loss of model performance. These potential extensions to the ICA approach provide considerable flexibility over the CCA approach in terms of model formulation, and will be reserved for future research.

[62] **Acknowledgments.** The authors wish to thank Luis Fernando Puerta Correa from the Empresas Publicas de Medellin (EEPPM) for providing hydrologic data. Funding for this research came from the Australian Research Council and the Sydney Catchment Authority. Their support for this work is gratefully acknowledged.

References

Aires, F., A. Chedin, and J. P. Nadal (1999), Analyse de series temporelles geophysiques et theorie de l'information: l'analyse en composants independants, *Geophysique externe, climat et environnement/External geophysics, climate and environment*, *328*, 569–575.

Aires, F., A. Chedin, and J. P. Nadal (2000), Independent component analysis of multivariate time series: Application to the tropical SST variability, *J. Geophys. Res.-Atmospheres*, *105*(D13), 17,437–17,455.

Aires, F., W. B. Rossow, and A. Chedin (2002), Rotation of EOFs by the independent component analysis: Toward a solution of the mixing problem in the decomposition of geophysical time series, *J. Atmos. Sci.*, *59*(1), 111–123.

Ashok, K., Z. Y. Guan, and T. Yamagata (2003), Influence of the Indian Ocean Dipole on the Australian winter rainfall, *Geophys. Res. Lett.*, *30*(15), 1821, doi:10.1029/2003GL017926.

Barnett, T. P., and R. W. Preisendorfer (1987), Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis, *Mon. Weather Rev.*, *115*, 1825–1850.

Barnston, A. G. (1994), Linear statistical short-term climate predictive skill in the northern hemisphere, *J. Clim.*, *7*, 1513–1564.

Barnston, A. G., and C. F. Ropelewski (1992), Prediction of ENSO episodes using canonical correlation analysis, *J. Clim.*, *5*, 1316–1345.

Basak, J., A. Sudarshan, D. Trivedi, and M. S. Santhanam (2004), Weather data mining using independent component analysis, *J. Mach. Learning Res.*, *5*, 239–253.

Casey, T. M. (1998), Assessment of a seasonal forecast model, *Aust. Meteorol. Mag.*, *47*, 103–111.

Chiew, F. H. S., and T. A. McMahon (2002), Modelling the impacts of climate change on Australian streamflow, *Hydrol. Processes*, *16*(6), 1235–1245.

Comon, P. (1994), Independent component analysis: A new concept?, *Signal Processing*, *36*, 287–314.

Drosdowsky, W., and L. E. Chambers (2001), Near-global sea surface temperature anomalies as predictors of Australian seasonal rainfall, *J. Clim.*, *14*, 1677–1687.

Filho, F. A. S., and U. Lall (2003), Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm, *Water Resour. Res.*, *39*(11), 1307, doi:10.1029/2002WR001373.

Goddard, L., et al. (2001), Current approaches to seasonal-to-interannual climate predictions, *Int. J. Climatol.*, *21*(9), 1111–1152.

Hamlet, A. F., D. Huppert, and D. P. Lettenmaier (2002), Economic value of long-lead streamflow forecasts for Columbia river hydropower, *J. Water Resour. Plann. Manage.*, *128*, 91–101.

Hastie, T., R. Tibshirani, and J. Friedman (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer Series in Statistics*, Springer, New York, 553 pp.

Herault, J., and C. Jutten (1986), Space or time adaptive signal processing by neural network models, in *Neural networks for computing: AIP conference proceedings American Institute for physics*, edited by J. S. Denker, New York.

Huang, J., H. M. Van Den Dool, and A. G. Barnston (1996), Long-lead seasonal temperature prediction using optimal climate normals, *J. Clim.*, *9*, 809–817.

Hurrell, J. W. (1995), Decadal trends in the North Atlantic Oscillation: Regional temperature and precipitation, *Science*, *269*, 676–679.

Hurrell, J. W., and H. Van Loon (1997), Decadal variations in climate associated with the NAO, *Clim. Change*, *36*, 301–326.

Hwang, S. O., J. K. E. Schemm, A. G. Barnston, and W. T. Kwon (2001), Long-lead seasonal forecast skill in far eastern Asia using canonical correlation analysis, *J. Clim.*, *14*, 3005–3016.

Hyvarinen, A. (1999), Survey on Independent Component Analysis, *Neural Comput. Surv.*, *2*, 94–128.

Hyvarinen, A., J. Karhunen, and E. Oja (2001), *Independent Component Analysis*, John Wiley and Sons, New York, 481 pp.

Ilin, A., H. Valpola, and E. Oja (2006), Exploratory analysis of climate data using source separation methods, *Neural Networks*, *19*, 155–167.

Kaplan, A., et al. (1998), Analyses of global sea surface temperature 1856–1991, *J. Geophys. Res. Oceans*, *103*(C9), 18,567–18,589.

Lee, T. W. (1998), *Independent Component Analysis - Theory and Applications*, Kluwer Academic Publishers, Boston.

Mason, S. J., and L. Goddard (2001), Probabilistic precipitation anomalies associated with ENSO, *Bull. Am. Meteorol. Soc.*, *82*(4), 619–638.

McBride, J. L., and N. Nicholls (1983), Seasonal relationships between Australian rainfall and the Southern Oscillation, *Mon. Weather Rev.*, *11*, 1998–2004.

Nicholls, N. (1989), Sea surface temperatures and Australian winter rainfall, *J. Clim.*, *2*, 965–973.

Oja, E. (Ed.) (2004), *Applications of Independent Component Analysis. Neural Information Processing - Lecture Notes in Computer Science, 3316*, Springer Berlin/Heidelberg, 1044–1051 pp.

Paegle, J. N., and K. C. Mo (2002), Linkages between summer rainfall variability over South America and sea surface temperature anomalies, *J. Clim.*, *15*(12), 19.

Richman, M. B. (1986), Rotation of principal components, *J. Climatol.*, *6*, 293–335.

Ropelewski, C. F., and M. S. Halpert (1996), Quantifying Southern Oscillation - Precipitation relationships, *J. Clim.*, *9*, 1043–1059.

Saji, N. H., and T. Yamagata (2003), Possible impacts of Indian Ocean Dipole mode events on global climate, *Clim. Res.*, *25*(2), 151–169.

Saji, N. H., B. N. Goswami, P. N. Vinayachandran, and T. Yamagata (1999), A dipole mode in the tropical Indian Ocean, *Nature*, *401*, 360–363.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice and Visualisation. Probability and Mathematical Statistics*, John Wiley & Sons Inc., New York, 317 pp.

Shabbar, A., and A. G. Barnston (1996), Skill of seasonal climate forecasts in Canada using canonical correlation analysis, *Mon. Weather Rev.*, *124*, 2370–2385.

Sharma, A. (2000a), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 - A strategy for system predictor identification, *J. Hydrol.*, *239*, 232–239.

Sharma, A. (2000b), Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3 - A non-parametric probabilistic forecast model, *J. Hydrol.*, *239*, 249–258.

Singhrattna, N., B. Rajagopalan, M. Clark, and K. K. Kumar (2005), Seasonal forecasting of Thailand summer monsoon rainfall, *Int. J. Climatol.*, *25*, 649–664.

Stone, R. C., G. L. Hammer, and T. Marcussen (1996), Prediction of global rainfall probabilities using phases of the Southern Oscillation Index, *Nature*, *384*, 252–255.

Storch, H. v., and F. W. Zwiers (2001), *Statistical Analysis in Climate Research*, Cambridge University Press.

Trenberth, K. E. (1997), The definition of El Nino, *Bull. Am. Meteorol. Soc.*, *78*, 2,771–2,777.

Verdon, D. C., A. M. Wyatt, A. S. Kiem, and S. W. Franks (2004), Multidecadal variability of rainfall and streamflow: Eastern Australia, *Water Resour. Res.*, *40*(10), W10201, doi:10.1029/2004WR003234.

Westra, S. P., C. Brown, U. Lall, and A. Sharma (2007), Modeling multivariable hydrological series: Principal component analysis or independent component analysis?, *Water Resour. Res.*, *43*, W06429, doi:10.1029/2007WR005617.

Wilks, D. S. (2006), *Statistical Methods in the Atmospheric Sciences. International Geophysics Series*, Elsevier, Amsterdam.

Woolddridge, S. A., M. F. Hutchinson, and J. D. Kalma (1999), Interpolation of rainfall data from raingauges and radar using thin plate smoothing splines, WATER99 Joint Congress. Institution of Engineers, Australia, Brisbane, Australia, pp. 263–268.

Yao, H., and A. Georgakakos (2001), Assessment of Folsom Lake response to historical and potential future climate scenarios - 2. Reservoir management, *J. Hydrol.*, *249*, 176–196.

C. Brown and U. Lall, Department of Earth and Environmental Engineering, Columbia University, New York, NY 10027, USA. (caseyb@iri.columbia.edu; ula2@columbia.edu)

A. Sharma, School of Civil and Environmental Engineering, The University of New South Wales, Sydney, NSW 2052, Australia. (a.sharma@unsw.edu.au)

S. Westra, Sinclair Knight Merz, 100 Christie Street, St Leonards, NSW, Australia. (swestra@skm.com.au)